Information-Theoretic Strategies for Quantifying Variability and Model-Reality Comparison in the Climate System

[1,2,3]**J. W. Larson**

[1]Mathematics and Computer Science Division, Argonne National Laboratory
9700 S. Cass Avenue, Argonne, IL 60439, USA
E-Mail: larson@mcs.anl.gov
[2] Computation Institute, University of Chicago
Chicago, IL, USA
[3] Department of Computer Science, The Australian National University
Canberra ACT 0200, Australia

Keywords: *Information Theory; Statistics; Climate Data Analysis*

Model-reality comparison can be viewed in a communications context. In this analogy, the observed "real" data are a sent message, and the model output are the received message. The model plays the role of a noisy channel over which the message is transmitted (Figure 1).

Information theory offers a way to assess literally the "information content" of any system, and offers a means for objective quantification of model-observational data fidelity. The Shannon entropy (SE) $H(X)$ is the measure of the amount of uncertainty, variability, or "surprise" present in a system variable X, while the Mutual Information (MI) $I(X,Y)$ measures the amount of shared information or redundancy between two variables X and Y. Information theory's roots lie in the analysis of communication of data across a noisy channel (Figure 1), and offer a scheme for quantifying how well a message X coming from a transmitter arrives as Y at the receiver. A more general information-theoretic measure of message degradation is the Kullback-Leibler Divergence (KLD), which quantifies insufficiency of agreement in the probatility desnity functions associated with X and Y. The ratio of MI to SE yields the amount of information shared by two datasets versus the information content of one alone. Alas, the aforementioned information-theoretic techniques work best for discrete rather than continuous systems. This is because evaluation of the Shannon Entropy (SE) for continuous systems–the differential entropy–does not constitute the continuum limit of the SE. Relative quantities such as the MI and KLD are always valid in the continuum case, and are the continuum limit of their discrete counterparts, but are just that–*relative*. This begs the question: Is there some way I can benchmark it against some continuum surrogate for the SE? Thus, one faces a choice when using information theory for model validation and intercomparison: (1) adopt coarse-graining strategies that are physically relevant, always aware that computed SE results re specific to a given discretisation; or (2) treat the data as continuous and use the MI combined with some benchmark quantity. In this paper, I adopt strategy (1), and restrict scope to a variable that has well-agreed-upon discretisations—total cloud cover, which by observational convention is frequently coarse-grained by oktas, tenths, or percent.

I review basic concepts from information theory. I put forward the notion that the SE is an alternative measure of climate variability, and evaluate it for reanalysis data and climate model output, producing global maps of the SE. I discuss how to structure sampling from two datasets to construct "messages" for use in information-theoretic model validation. I derive from the SE and MI a pair of fidelity ratios for assessing model-reality fidelity, and evaluate them for total cloud amount. I apply a modified KLD to asses model-reality agreement for local temporally-sampled total cloud, and explain the relative strictness of the KLD- and MI-based validation standards. I conclude with a a roadmap for analysing and validating the informatics of climate.

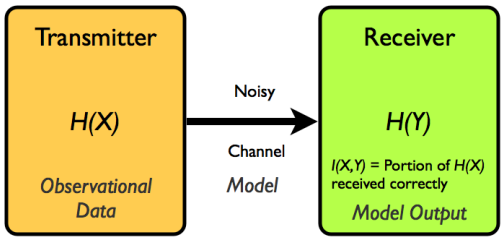

Figure 1. Communications system with source producing $H(X)$ and receiver seeing $H(Y)$. Amount of data communicated correctly from source to receiver is the mutual information $I(X,Y)$.

1 INTRODUCTION

Climate model output evaluation remains an area of active research. Many researchers rely on comparison of statistical moments such as the mean and variance, or on correlation analysis. Moment-based statistical tests such as the t- and F-tests rely on an assumption of normality of the underlying population. Correlation analysis between variables is appropriate under the assumptions of normality, linearity, and homoskedasticity. Information theory provides an attractive approach to higher-order statistical analysis that avoids the assumptions associated with correlation analysis and moment-based hypothesis tests. The strategy in information theory is based on the underlying probability density either for a finite set of states for a discrete variable, or for a probability density function for a continuous variable.

In this study, I explore the idea of Shannon entropy as an indicator of climate variability. I also present two new quantities for assessing model-reality fidelity that are based on Shannon entropy and mutual information. I define a procedure for computing these quantities and estimating associated uncertainties. I find these "fidelity ratios" impose a very high standard of model-reality fidelity that is hard to meet for a typical climate model. I employ a more lenient standard for model probability density comparison—the Kullback-Leibler divergence—and explain how the low fidelity ratios result in some cases due to poor agreement between the model's and reanalysis' respective probability densities.

This is not the first use of information-theoretic quantities in climatology. Bagrov first introduced a "similarity index" for meteorological model-reality comparison that assumed underlying continuous normal distributions (Bagrov [1963]). Much work has been done on the use of mutual information as an indicator of predictability (DelSole and Tippett [2006]). Muual information has also been employed to study relationships between climate variables(Knuth et al. [2005]). Relative entropy has been used to validate global distributions of surface temperature (Shukla et al. [2006]). To my knowledge, this is the first use of information theory to express climate variability, and to present geographic distributions of H, MI-based fidelity ratios, and the KLD.

In this study I use total cloud cover to illustrate informatic climate variability. Total cloud cover has the advantage of having standardised discretisations amenable to discrete-variable informatics, and is of climatological significance because it is an integrated diagnostic of parameterisations of atmospheric column physics, feeds back into atmospheric radiative transfer, and is a variable for which widespread observations exist.

2 INFORMATION THEORY

Here we review key concepts from information theory and define terms used in the rest of this paper. Further details may be found in standard textbooks (Cover and Thomas [2006]; Reza [1994])..

Consider a discrete variable X that can have any of N possible values; $X \in \{x_1, \ldots, x_N\}$. The probability of observing each value $X = x_i$ is $0 \leq p(x_i) \leq 1$, subject to the constraint $\sum_{i=1}^{N} p(x_i) = 1$. The *Shannon Entropy* (SE) or $H(X)$ is defined as

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log [p(x_i)]. \tag{1}$$

The units of H depend on the base of the logarithm; for base 2 H is measured in *bits*, for natural base e H is measured in *nats*. If X is the set of values seen in a signal, then $H(X)$ is the *amount of information in the signal*. Note also that the SE is nonnegative and finite.

Consider two discrete variables $X \in \{x_1, \ldots, x_N\}$ and $Y \in \{y_1, \ldots, y_M\}$ defined with respective probabilities $\{p(x_1), \ldots, p(x_N)\}$ and $\{p(y_1), \ldots, p(y_M)\}$, subject to the above normalisation and nonnegativity constraints used to define the SE. The probability of seeing the combination (x_i, y_j) is the *joint probability* $0 \geq p(x_i, y_j) \leq 1$, and subject to the normality constraint $\sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) = 1$. The *joint entropy* $H(X, Y)$ measures the combined information content of X and Y, and is defined as

$$H(X, Y) = \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) \log [p(x_i, y_j)]. \tag{2}$$

If the variables X and Y are statistically independent, then the joint entropy $H(X,Y)$ is the sum of the SEs $H(X)$ and $H(Y)$. If X and Y are somehow related and share information, then

$$H(X,Y) = H(X) + H(Y) - I(X;Y), \tag{3}$$

where $I(X;Y)$ is the *transinformation* or *mutual information* (MI)

$$I(X;Y) = \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i,y_j)\log\left[\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right]. \tag{4}$$

The units for the MI are dictated by base for the logarithm in (4), just as the units are for the SE in (3). The MI is symmetric; that is $I(X;Y) = I(Y;X)$. If the variables X and Y constitute identical signals, then $H(X) = H(Y) = I(X;Y)$. The MI satisfies the properties $0 \le I(X,Y) \le H(X)$ and $0 \le I(X,Y) \le H(Y)$. The fidelity of transmitting a signal X and receiving Y can be quantified using the ***fidelity ratios***

$$F_{YX} = \frac{I(X;Y)}{H(X)} \qquad \text{and} \qquad F_{XY} = \frac{I(X;Y)}{H(Y)} \tag{5}$$

F_{YX} is the fraction of information present in signal X that was successfully transmitted to Y; note that $0 \le F_{YX} \le 1$. F_{XY} is the fraction of information present in signal Y that was successfully received from X; note that $0 \le F_{XY} \le 1$. In the case of perfect transmission of source X to receiver Y, $F_{XY} = F_{YX} = 1$.

As we will see in Section 4, the mutual information is a high standard of quality for dataset intercomparison. Another approach is to ask: How well do two probability densities that share a common partitioning scheme agree? The *Kullback-Leibler Divergence* (KLD) or *relative entropy* is an information-theoretic standard for judging how well two probability densities based on a common partitioning scheme agree. Suppose for some variable $X \in \{x_1,\ldots,x_N\}$ we have two candidate probability densities $p(X)$ and $q(X)$, which may be viewed as the "true" and "modeled" densities, respectively. The KLD D_{KL} is defined as

$$D_{KL}(p \parallel q) = \sum_{i=1}^{N} p(x_i)\log\left[\frac{p(x_i)}{q(x_i)}\right]. \tag{6}$$

The KLD is not symmetric; that is, $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$. The units for the KLD are defined the same way as for the SE and MI. The KLD is sometimes called the *Kullback-Leibler gain* or *information gain* required to represent $p(X)$ given $q(X)$. the "true" and "observed" distributions for X, respectively. On average one needs $D_{KL}(p \parallel q)$ extra bits of information per symbol to represent $p(x)$ using $q(x)$ as a starting point. The KLD is nonnegative. If there is perfect agreement between $p(X)$ and $q(X)$, $D_{KL} = 0$. There is no upper bound for values of D_{KL}; for example, singularities can arise in (6) if $q(x_i) = 0$ and $p(x_i) \neq 0$, leading to infinite KLD..

For a continuous variable $X \in (-\infty,\infty)$, it is possible to define a *differential entropy* (DE) H(X) for $x \in (-\infty,\infty)$ by replacing the marginal probabilities $p(x_i)$ with a continuous probability density function $p(x)$, and replacing the summation over state index i in (1) with an integral with respect to x. It is tempting to think that the DE is the continuum limit of the SE (3); alas, it is not a valid measure of information content because the integral in the definition of the DE is sensitive to the bin widths dx, and because it is possible for $p(x) > 1$ for some values of x, thus making it possible to have infinite or negative values of the DE. Furthermore, the values of the DE are not invariant under coordinate transformations. Two information-theoretic quantities are, however valid in the continuum limit: The mutual information and the Kullback-Leibler divergence. In this study, the scope is restricted to discretised variables whose quantisation arises from meteorological observation conventions.

3 DATA AND ANALYSIS

The "reality" data used in this study are the National Center for Environmental Prediction / Department of Energy Reanalysis 2 dataset (NCEP-2; Kanamitsu et al. [2002]) that cover the period January 1979-December

2008. Monthly averages are drawn from this dataset, which can be downloaded from the NCEP-2 Web site (National Oceanagraphic and Atmospheric Administration Earth System Research Laboratory [2009]). The data reside on a T62 Gaussian grid with 192 longitudes and 96 latitudes. There are 360 monthly averages in the sample at each grid location. The NCEP-2 total cloud amount data ("tcdc") is used in this study, and have values in percent cloud cover ranging from zero to 100 percent. The "model" data are from a 500-year control run of the Community Climate System Model (CCSM3; Collins et al. [2006]). Monthly averages were drawn from the repository of CCSM3 model integration output data maintained by the Earth System Grid (United States Department of Energy and University Corporation for Atmospheric Research [2009]). The data reside on a T85 Gaussian Grid comprising 256 longitudes and 128 latitudes. There are 6000 monthly averages in the sample at each grid location. The CCSM3 total cloud amount data ("CLDTOT") is used in this study, and have values ranging from zero to 1. For the SE calculations in this study, the data were used on their respective grids. For the MI and KLD calculations, the CCSM3 data were interpolated from their T85 grid the the NCEP-2 T62 grid using an inverse-great-circle-distance weighted scheme that is valid assuming the geoid is a sphere.

Cloud amounts in the reanalysis and model data were coarse-grained into oktas, tenths, and percent, thus avoiding problems associated with the DE. Data values are mapped into the interval $[0, 1]$. The data are then multiplied by 8, 10, or 100 for oktas, tenths, or percent, respectively. A class value is assigned by rounding to the nearest integer to the data value. Thus, 9, 11, and 101 classes result from coarse-graining by oktas, tenths, and percent, respectively. The data are organised as time series of global geographic distributions; that is, they are three-dimensional datasets with dimensions longitude, latitude, time). For this study, time series for fixed values of longitude and latitude are used as the samples from which SE, MI, and KLD are computed. Thus, maps of of these quantities may be drawn to illustrate the geographic distribution of entropy and other information statistics.

Calculation of the SE and MI from equations (3) and (4) are straightforward; any instances in which $p(x_i) = 0$ in (3) and $p(x_i, y_j) = 0$ in (4) provide zero contribution to the SE and MI, respectively. Fidelity ratios F_{YX} and F_{XY} are computed using (5). Singularities can arise in the KLD calculation using (6). For this study, singularities in the KLD calculation are avoided through addition of a small observability threshold term ι to each value of $q(x_i)$, and subsequent renormalisation by division by $1 + N\iota$, where N is the number of classes in the coarse-graining scheme. The observability threshold was chosen to be $\iota = \frac{1}{2N_S}$, where N_S is the number of time samples (6000 for the CCSM3 data). Thus, large—rather than infinite—KLD values result where class i is observed for $p(x_i)$ but not for $q(x_i)$.

A sliding window sampling scheme is used to estimate uncertainties in the SE, MI, fidelity ratios, and KLD. For the SE calculations, a 20-year window is used to compute H, and the window is advanced one year, removing the first year from the sample, and introducing a new year at its end. For the NCEP-2 data and CCSM3 data this results in 11 and 481 samples for their respective SE calculations. The mean $\langle H \rangle$ and standard deviation σ_H are computed from the ensemble of resulting SE values. For the MI and KLD calculations, all 30 years of the NCEP-2 data and a sliding 30-year window of CCSM3 data are used, resulting in an ensemble of 471 values of the MI and KLD. Ensemble averages and standard deviations are subsequently computed for the MI, fidelity ratios, and KLD.

4 RESULTS

SE for total cloud cover discretised by oktas and percent from a thirty-year sample of NCEP-2 and CCSM3 data are shown in Figure 2. Fields of H for total cloud discretised by tenths for a twenty-year sliding window sample of NCEP-2 and CCSM3 with their associated uncertainties σ_H are presented in Figure 3. The values of the SE are quite sensitive to the number of classes, but the overall spatial structure of the SE fields is preserved. In both NCEP-2 and CCSM3 data, relatively high SE values are associated with the tropics, particularly in monsoon regions. The lowest values of SE lie in a band over the Southern Ocean centered at approximately $50°$S. CCSM3 data have much more widespread high SE regions over land than NCEP-2, and have regions of high entropy in Western Asia and the US Pacific Northwest that are not present in the reanalyses. These high entropy regions are associated with relatively flat probability densities for total cloud, and in this sense indicate greater variability. The associated uncertainties σ_H shown in the right panels in Figure 3 are small, at the most on the order of $\leq 1\%$.

The fidelity ratio fields derived from the MI and SE for CCSM3 vs. NCEP-2 total cloud are shown in Figure 4. Note that worldwide these values are low, with $F_{YX} \leq 35\%$. Over some areas of poor agrement (e.g., the Southern Ocean), there is considerable noise ($\sigma_{F_{YX}}/F_{YX} \approx 10\%$ of its raw value) in the results, indicative of variability in the ordering of the tenths classes. The areas of best agreement are over land masses associated with monsoons and in other regions such as the US Pacific Northwest, the Middle East, and west-central Asia. Signal-to-noise ratios

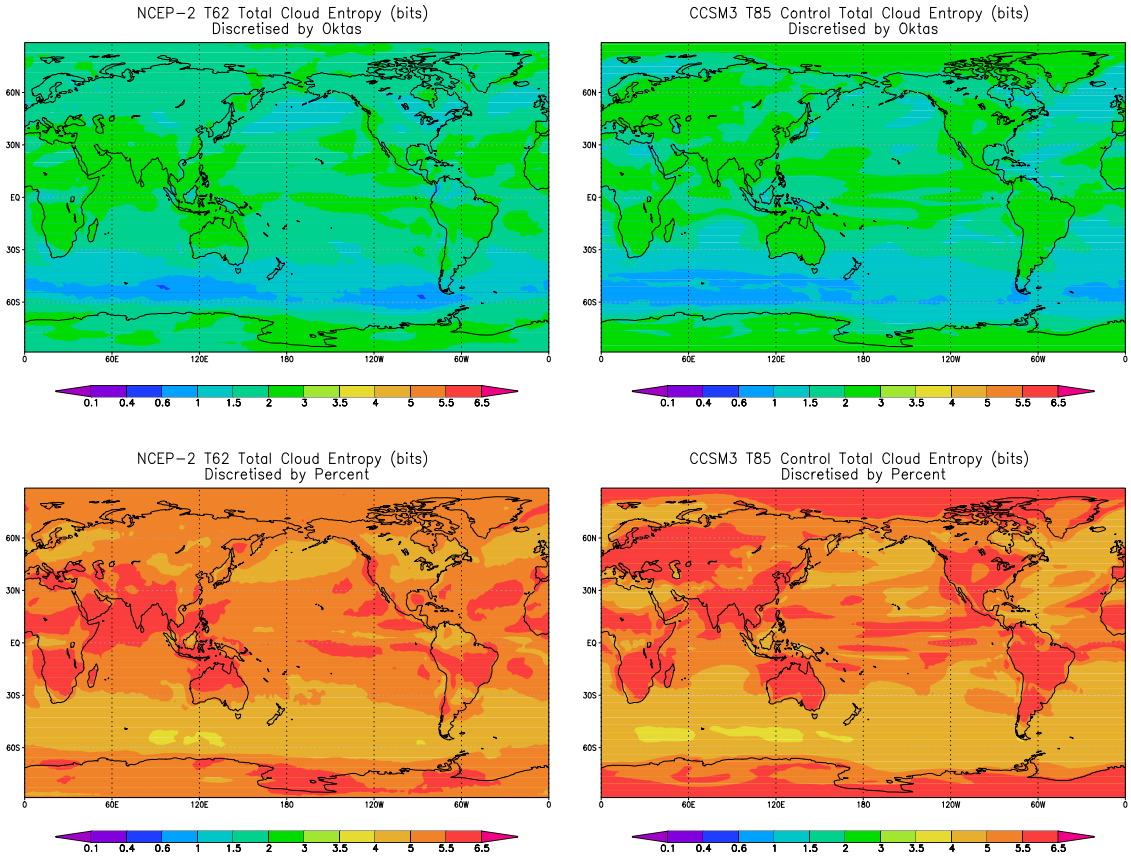

Figure 2. Shannon entropy for total cloud for CCSM3 and NCEP-2 using various discretization strategies.

for these regions are high in the monsoonal areas, but low in the other areas with large F_{YX}. From this MI-based analysis, the temporal structure of the occurrence of tenths classes agrees poorly. This is in part due to interannual variability, but may have other causes stemming from model bias.

A model bias cause of poor performance in the fidelity ratio metrics shown in Figure 4 may be underrepresented or absent classes in the probability density for CCSM3 total cloud discretised by tenths. The KLD offers a scheme for testing probability density quality, and the KLD field total cloud is shown in Figure 5, with the NCEP-2 and CCSM3 cloud probability densities playing the roles of $p(x_i)$ and $q(x_i)$ in equation (6), respectively. For much of the world, low KLD values indicate that the probability densities associated with CCSM3 total cloud agree well with with their reanalysis counterparts, particularly over land masses. Notable exceptions are polar regions over land, bands over ocean at $30°$N and $30°$S, an equatorial band over ocean stretching from the Eastern Pacific and across the Atlantic Oceans, and a region in off the west coast of South America. The associated uncertainties in the KLD values σ_{KLD} are bounded above by 10%. Some of the higher KLD values in these regions are caused by absence of cloud amount classes in the CCSM3 probability densities, and are the singular terms in the KLD mentioned in Section 2. Of particular interest are the regions of good agreement in probability densities over land that score poorly in terms of MI, for example parts of Asia, Australia and the Americas. In these areas, CCSM3 is reproducing the probability density well, but not its annual and interannual ordering of cloud amount classes.

5 CONCLUSIONS AND FUTURE WORK

An information-theoretic approach to climate variability has been presented and its utility in analysing total cloud amount variability and model-reanalysis comparison have been demonstrated. This is the first stage in a much larger plan to study the overall informatics of the climate system. Most climate variables are quantities that have no standard discretisations, which will require a rigourous strategy for coarse-graining climate data such as the sample-based optimal binning strategy for pdf estimation (Knuth [2006]). Future work will proceed on a number of fronts including the spatial informatics, informatic relationships between multiple climate variables, and how the informatic structure of the climate system may change due to anthropogenically-induced global warming.

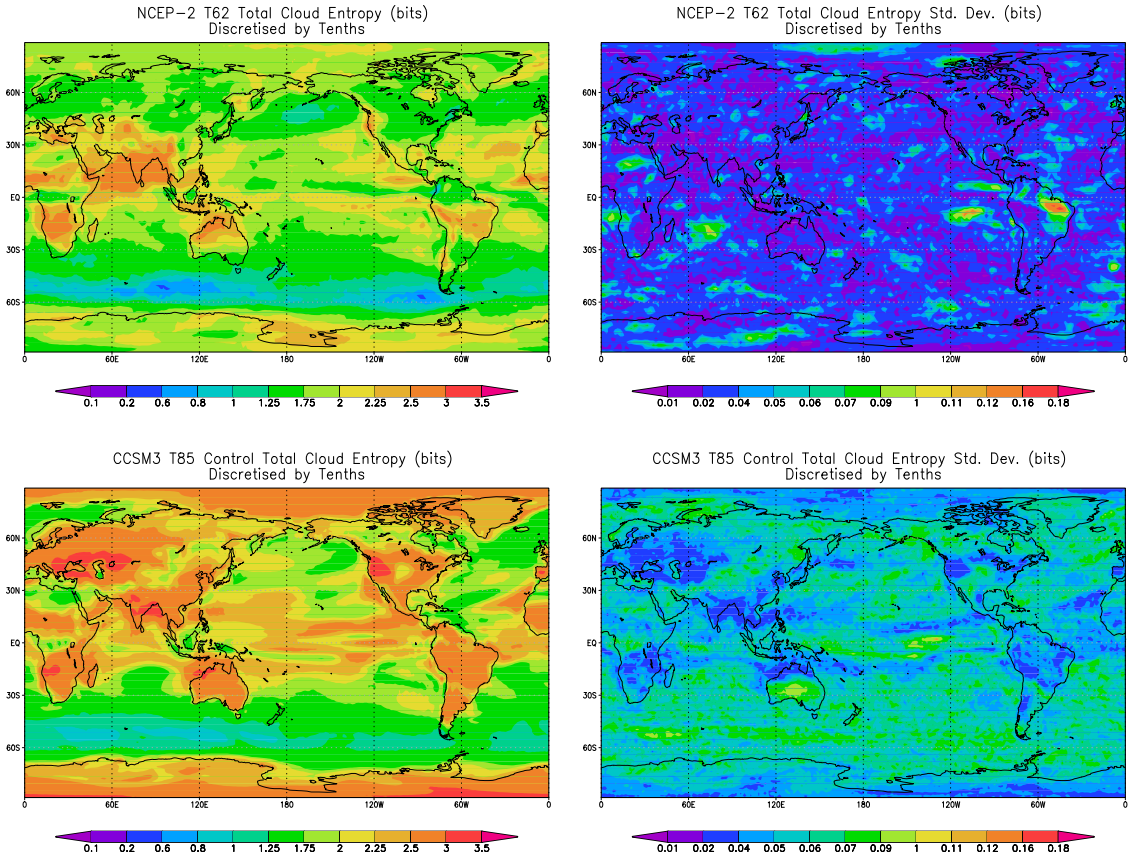

Figure 3. Shannon entropy H for total cloud for CCSM3 and NCEP-2 with sample standard deviation σ_H.

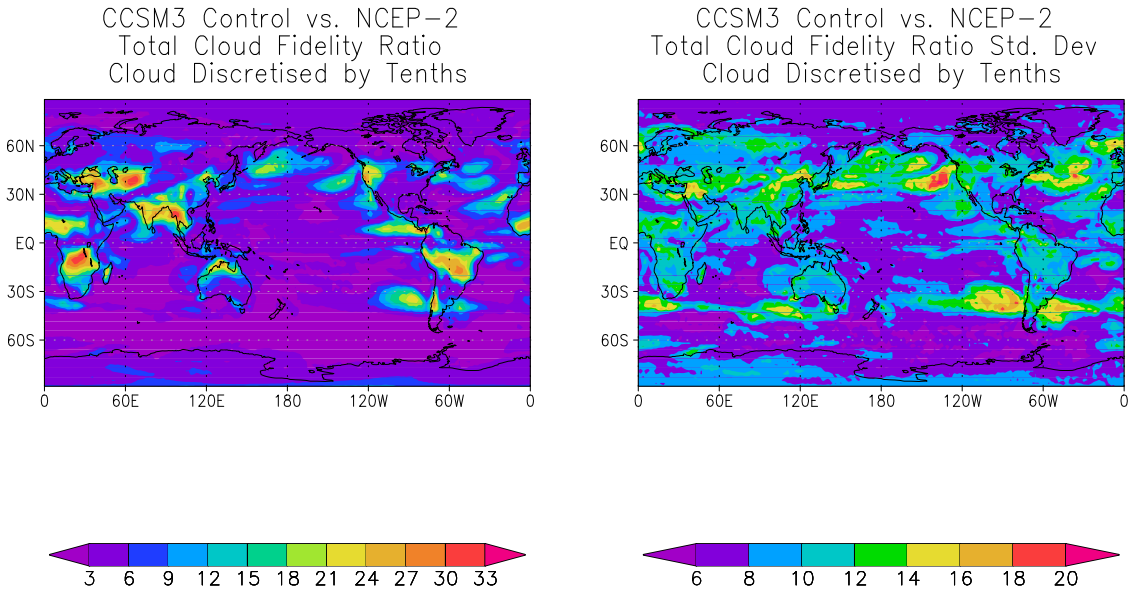

Figure 4. Fidelity ratio F_{YX} for total cloud for CCSM3 and NCEP-2 with sample standard deviation σ_I. Units for colourscale for F_{YX} and $\sigma_{F_{YX}}$ are in percent and thousandths, respectively.

ACKNOWLEDGMENTS

This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy (DOE), under Contract DE-AC02-06CH11357 and by the Climate Change Prediction

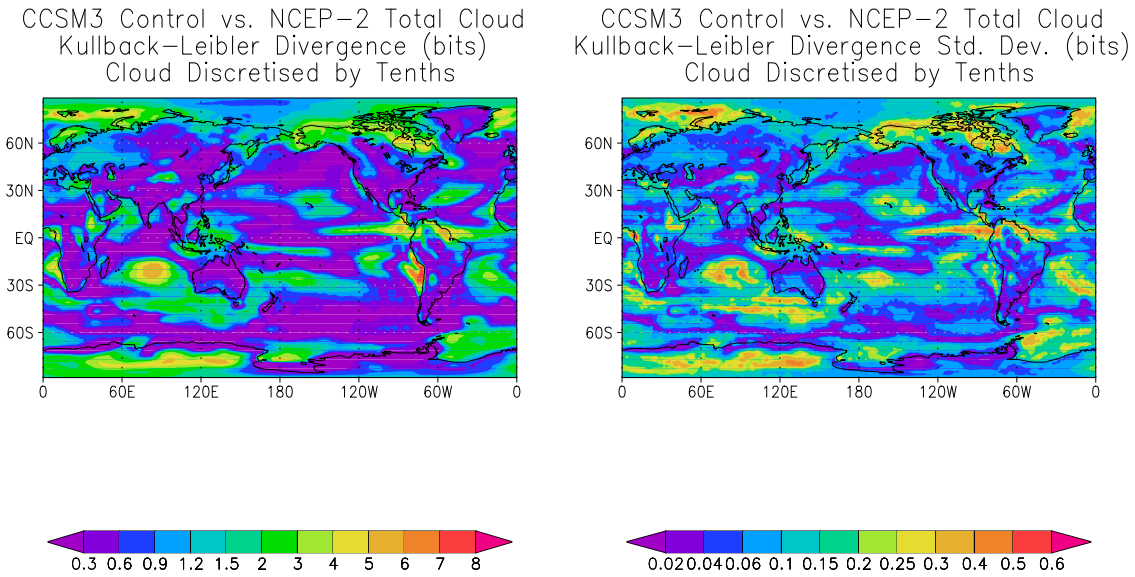

Figure 5. Kullback-Leibler Divergence D_{KL} for total cloud for CCSM3 and NCEP-2 with sample standard deviation σ_{KLD}.

Program. I thank the Department of Theoretical Physics of the Research School of Physical Sciences and Engineering for hosting me as a visiting fellow.

REFERENCES

Bagrov, N. A. Statistical entropy as an indicator of simarity or difference of meteorological fields. *Meteorologiya i Gidrologiya*, (1), 1963.

Collins, W. D., C. M. Bitz, M. L. Blackmon, G. B. Bonan, C. S. Bretherton, J. A. Carton, P. Chang, S. C. Doney, J. J. Hack, T. B. Henderson, J. T. Kiehl, W. G. Large, D. S. McKenna, B. D. Santer, and R. D. Smith. The Community Climate System Model: CCSM3. *Journal of Climate*, 19(11):2122–2143, 2006.

Cover, T. M. and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.

DelSole, T. and M. Tippett. Predictability: Recent insights from information theory. *Reviews of Geophysics*, 45: RG4002, 2006.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S.-K. Yang, J. J. Hnilo, M. Fiorino, and G. L. Potter. NCEP-DOE AMIP-II reanalysis (R-2). *Bulletin of the American Meteorological Society*, 83(11):1631–1643, 2002.

Knuth, K. H. Optimal data-based binning for histograms. `http://arxiv.org/abs/physics/0605197`, 2006.

Knuth, K. H., A. Gotera, C. T. Curry, K. A. Huyser, K. R. Wheeler, and W. B. Rossow. Revealing relationships among relevant climate variables using information theory. In *Proceedings of the Earth System Technology Conference (ESTO 2005)*, Lecture Notes in Computer Science, 2005.

National Oceanagraphic and Atmospheric Administration Earth System Research Laboratory. NCEP-DOE-AMIP-II Reanalysis. `http://www.cdc.noaa.gov/data/gridded/data.ncep.reanalysis2.html`, 2009.

Reza, F. M. *An Introduction to Information Theory*. Dover, New York, 1994.

Shukla, J., T. DelSole, M. Fennessy, J. Kinter, and D. Paolino. Climate model fidelity and projections of climate change. *Geophysical Research Letters*, 33:L07702, 2006.

United States Department of Energy and University Corporation for Atmospheric Research. Earth System Grid Web site. `http://earthsystemgrid.org`, 2009.

LICENSE